# THE RESEARCH ON THE OPTIMUM OVERSAMPLING RATIO OF DEFAULT EVENT - THE CASE OF PERSONAL CONSUMER LOANS DEFAULT PREDICTIVE MODEL

TE-HSIN LIANG

*Department of Statistics and Information Science,*
*Fu Jen Catholic University, Taipei, Taiwan*
*stat1013@mail.fju.edu.tw*

JIAN-BANG LIN

*Ph. D. Student of Business Administration,*
*Fu Jen Catholic University, Taipei, Taiwan*
*bang.lin@msa.hinet.net*

SHIH-TING TSAI

*Graduate Institute of Applied Statistic,*
*Fu Jen Catholic University, Taipei, Taiwan*
*atong428@gmail.com*

The New Basel Capital Accord will allow banks to determine their regulatory capital requirements using the Probability of Default (PD) on personal consumer loans. Within the new revisions, banks are allowed to determine their capital charges using collected data to compute the inherent credit risk of each borrower. To do this, the prediction of the PD has its urgency and necessity. Many researches proposed that there were different model effects under various oversampling ratio of default event, but rarely discussed the optimum oversampling ratio of default event corresponding to the diverse management strategies of bank. The purpose of this research is to find the optimum oversampling ratio of default event in the most effective predictive default model for the diverse management strategies. If the bank's strategy focuses on the accuracy rate or the precision rate, the result shows that the optimum oversampling ratio of default event should be as similar as population dataset. The oversampling ratio of default to non-default event 3:1 will be the best, when the bank considers the recall rate is the most important issue. Another, we propose that the oversampling ratio 1:2 is the best when the bank expects the true negative rate to be good.

## 1. Introduction

Through receiving surplus bankrolls from depositors and loaning them to accommodators, a bank plays the important role of mediator between depositors

and accommodators. If the bank is well operated, it will raise the circulation of bankrolls and create lots of profit. Contrary, once the bank does not do risk-management well, it influences not only itself and depositors but also the financial market, even the whole country. So, the risk-management is one of the most important links in successful bank management.

In order to gain more profit under the highly competitive finance market, many banks offer personal consumer loans. Those products of personal consumer loans emphasize small amount, without guarantee, low interest rate, and easy applying. The auditing of personal consumer loans is looser than other loan products; this will cause the decrease of loans quality and the increase of "Non-performing loan ratio (Default[a])", and might result in another serious financial crisis. That is, to construct a systematic and effective default predictive model is an important and urgent subject for banks.

According to the literature, many researches proposed that there were different model effects under various oversampling ratio of default event. However, when building the default predictive model, the optimum oversampling ratio of default event corresponding to the diverse management strategies of bank was rarely discussed in the past. So, the purpose of this research is to find the optimum oversampling ratio of default event in the most effective default predictive model according to the diverse management strategies.

## 2. Literature review

Credit risk had been studied early in the 1950s. Ohlson [7] pioneered Logistic Regression (LR) model to construct credit risk model, and lots of comparison studies have affirmed that LR model is more powerful [1] [4] [5] [8]. So, this research applies LR to construct the default predictive model.

In the past, most scholars used the oversampling ratio of default to non-default event 1:1 to construct credit risk model [1] [3] [4] [6] [7] [10]. Zavgren [14] suggested using the oversampling ratio of default to non-default event 1:1 in constructing predictive model, but did not discuss systematically in her research. Shi [13] and Liang et al. [11] proposed that the oversampling ratio of default to non-default event 1:3 is the best for Mainland China and Taiwan. However, their conclusions were made by considering either the lowest error rate or the highest accuracy rate among all cases of default and non-default events. But different departments or strategies of bank will pay attention to diverse issue; while the department of risk control may care about the accurate

[a] The obligor is past due for more than 90 days.

rate of predicted default events, the department of business may focus on the accurate rate of predicted non-default events. So, this research intends to find the optimum oversampling ratio of default event in constructing the default predictive model according to diverse management strategies of bank.

## 3. Methods

### 3.1. *Variable selection*

In this research, with the significance level at 0.05, t test (for continuous variables) and Chi-square test (for categorical variables) are used to select the independent variables which significantly influence default. To solve the multi-colinearity problem, among independent variables of highly correlative, the most significant one to default will be chosen by comparing their values of Eta-square or Cramer's V.

### 3.2. *Data partition and Oversampling*

When constructing the default predictive model, the dependent variable is usually a rare event. Berry and Linoff [9] suggested that applying oversampling to raise the ratio of rare events in the sample will make it easier to construct a better predictive model. The default rate in this research is 16.30% and is rare contrasting to non-default event. This research administers 7 oversampling ratios of the default and non-default event including 3:1, 2:1, 1:1, 1:2, 1:3, 1:4, and 1:5 when constructing the default predictive model. In order to confirm the validity of the default predictive model, the whole dataset is partitioned into 80% (training dataset) and 20% (testing dataset). From training dataset, twenty samples are re-sampling for each oversampling ratio. In other words, the default predictive models are reconstructed 20 times for each oversampling ratio. That is, this research uses 140 samples in constructing the default predictive models. These 140 samples are tested for their Homogeneity to the population, respectively in default and gender, by the Chi-square test with the significance level at 0.05.

### 3.3. *LR model*

LR is similar to general linear regression, but its dependent variable is binary or multinomial. LR is better than the Multiple Discriminant Analysis, because it allows to use both categorical and continuous independent variables when constructing the model [2]. In addition, the LR model will not only predict classification but also calculate the probability of default for each case. LR uses

the Maximum-likelihood method to estimate parameters [12]. It is represented by the following Eq. (1):

$$\text{logit}\left[\pi(\text{x})\right] = \log\left(\frac{\pi(\text{x})}{1-\pi(\text{x})}\right) = \beta_0 + \beta_1 \text{X}_1 + \ldots + \beta_k \text{X}_k \quad (1)$$

where $\pi(\text{x})$ is the probability of event, $0 \leqq \pi(\text{x}) \leqq 1$

### 3.4. *Confusion matrix*

This research applies one of the most common evaluate procedures, confusion matrix (see Table 3.1), to find the effective of models. Four evaluating indexes are used: the Accuracy Rate (AR) shows the rate that events (default and non-default) are predicted correctly in all cases; the True Negative Rate (TNR) shows the rate that the non-default events are predicted correctly in all true non-default cases; the Recall Rate (RR) shows the rate that default events are predicted correctly in all the true default cases; and the Precision Rate (PR) shows the rate that default events are predicted correctly in all predicted default cases. The higher the indexes are, the better the model is. The formulas are as Eq. (2):

$$AR = \frac{(A+D)}{(A+B+C+D)} \times 100\,\%, \qquad TNR = \frac{D}{(B+D)} \times 100\,\%$$
$$RR = \frac{A}{(A+C)} \times 100\,\% \quad , \qquad PR = \frac{A}{(A+B)} \times 100\,\% \quad (2)$$

Table 3.1 Confusion matrix

| Prediction \ True | Default | Non-default |
|---|---|---|
| Default | A | B |
| Non-default | C | D |

### 3.5. *Effectiveness index*

In order to compare the effectiveness of the default predictive model among the 7 oversampling ratios, this research calculated AR, TNR, RR and PR on 20 different cut-points of the predicted default probabilities. We choose the percentiles of 7, 8, 9, 10, 11…, 23, 24, 25 and 26 because of their nearing to the default rate of population 16.30%. We used "Superior Rate (SR)" which was revised from the Liang et al. [11] to evaluate the effectiveness of the default predictive model on different oversampling ratios. First, the $I_{ij}$ is defined as the value of the evaluating index, including AR, TNR, RR and PR, at the $i^{th}$ cutpoint

and the $j^{th}$ oversampling ratio of default predictive model. The $S_{ij}$ and $SR$ are given by Eq. (3) and (4).

$S_{ij}$=1 if $I_{ij}$ is the highest value at the $i^{th}$ cut-point, otherwise $S_{ij}$=0. (3)

$$SR_j = \frac{\sum_{i=1}^{m} S_{ij}}{m} \quad (4)$$

where $m$ is the number of cut-points, in this research $m$=20.

$SR_j$ is the Superior Rate at the $j^{th}$ oversampling ratio. The higher $SR_j$ means the better predicting.

## 4. Results

### 4.1. *Data structure*

The data in this research are 10,470 cases of the personal consumer loans approved by a local bank in Taiwan between 2003/06~2004/06. There are 1,707 defaults (16.30%) and 8,763 non-defaults (83.70%). After consulting banking professionals, we deleted independent variables which have no significantly influence to default. Fifty-two significant independent variables are chosen, which are classified into two groups including 24 variables from bank itself and 28 independent variables from JCIC (the Joint Credit Information Center (JCIC) in Taiwan).

### 4.2. *Variables selection*

As mentioned in sec. 3, t test, Chi-square test, Eta-square and Cramer's V are applied to select the independent variables which have significantly influencing to default. Finally, 28 significant independent variables are selected (see Table 4.1).

### 4.3. *The optimum oversampling ratio*

Using twenty re-samples, 20 default predictive models for each oversampling ratio are constructed. Then, the average of the predicted default probabilities of 20 default predictive models are calculated for each oversampling ratio model and 7 ensemble default predictive results are found. The effectiveness comparisons among ensemble results of 7 oversampling ratio models are done by using confusion matrix and index of SR.

Table 4.2 shows that the bank can adopt different oversampling ratio according to different management strategies. If the bank focuses on the AR or the PR, 1:5 is the optimum oversampling ratio of default event in constructing

effective default predictive model. The oversampling ratio 1:5 is similar to the default rate of population dataset. If bank focuses on the AR or the PR, the oversampling technique might not be needed. If the bank considers the RR is the most important to the loans, the oversampling ratio of default to non-default 3:1 is suggested to be the best. Another, we propose that the oversampling ratio of default to non-default 1:2 is the best when the bank expects TNR of the default predictive model to be good.

Table 4.1 The table of significant independent variables

| type | Variable | p-value | MOA* | Variable | p-value | MOA* |
|---|---|---|---|---|---|---|
| Cont. | ●Apply_interest | 0.000 | 0.018 | ●Job_years | 0.035 | 0.010 |
| | ●Min._payment_amt. | 0.000 | 0.702 | ●Credit_limit_amt. | 0.000 | 0.174 |
| | ●Cashcard_loan_balance | 0.000 | 0.145 | ●Last_3_mon._query_cnt. | 0.000 | 0.062 |
| | ●Credit_card_revolver_rate | 0.000 | 0.055 | ●Last_other banks_ query_days | 0.000 | 0.028 |
| | ●First credit card_age | 0.000 | 0.027 | ●Rate of last_n_mon._deferred_cnt. | 0.012 | 0.051 |
| Cate. | ●Credit card_exception | 0.000 | 0.341 | ●Gender | 0.000 | 0.133 |
| | ●Education | 0.000 | 0.117 | ●Company_add. | 0.000 | 0.097 |
| | ●Located_area | 0.000 | 0.090 | ●Age | 0.000 | 0.067 |
| | ●Case_source | 0.000 | 0.065 | ●Branch | 0.000 | 0.056 |
| | ●Job_position | 0.001 | 0.055 | ●Occupation | 0.000 | 0.050 |
| | ●Mortgage | 0.003 | 0.029 | ●Car_loan | 0.005 | 0.029 |
| | ●Building_type | 0.006 | 0.043 | ●Annual _income | 0.020 | 0.020 |
| | ●Cellphnoe | 0.023 | 0.023 | ●Permanent_phone | 0.048 | 0.020 |
| | ●Payment_code | 0.000 | 0.059 | ●Last_8_mon._min._unpay_cnt ≧ 4 | 0.000 | 0.044 |

Note *: MOA means measures of association, and they are Eta-square and Cramer's V in this research.

Table 4.2 The table of SR at 4 indices (%)

| index \ ratio | 1：1 | 1：2 | 1：3 | 1：4 | 1：5 | 2：1 | 3：1 |
|---|---|---|---|---|---|---|---|
| AR | 0.00 | 25.00 | 5.00 | 20.00 | 40.00 | 0.00 | 0.00 |
| TNR | 0.00 | 35.00 | 0.00 | 20.00 | 25.00 | 0.00 | 0.00 |
| RR | 10.00 | 0.00 | 20.00 | 15.00 | 10.00 | 10.00 | 45.00 |
| PR | 5.00 | 15.00 | 10.00 | 15.00 | 45.00 | 0.00 | 0.00 |

## 5. Conclusion

As the financial environment is ever-so-competitive, every bank wants to acquire the most profit and the least loss. This research tried to find the optimum oversampling ratio to construct the effective default predictive model according to different management strategies for Taiwan local banks. The conclusion as following can be the reference for constructing default predictive model in the future:

1. If the bank focuses on the AR or the PR, the oversampling might not be needed; that is, when constructing effective default predictive model, we

suggest that the sample ratio of default should be similar to the population dataset.

2. If the bank focuses on the RR, the optimum oversampling ratio of default to non-default event is 3:1 when constructing effective default predictive model.
3. If the bank focuses on the TNR, the optimum oversampling ratio of default to non-default event is 1:2.

**References**


1. A. W. Lo, "Logit Versus Discriminant Analysis- A Specification Test and Application to Corporate Bankruptcies," *Journal of Econometrics*, Vol. 31, issue 2, pp. 151-178 (1986).
2. D. W. Hosmer and S. Lemeshow, "Applied logistic regression," New York: John Wiley and Sons (2000).
3. E. I. Altman, "Financial Ratios, Discriminant Analysis and the Prediction of Corporate Bankruptcy," *The Journal of Finance*, pp. 589-609 (1968).
4. Espahibodi, "Identification of Problem Bank and Binary Choice Models," *Journal of Banking and Finance*, Vol. 15, Issue 1, pp. 53-71 (1991).
5. F. E. Harrel and K. L. Lee, "A Comparison of the Discrimination of Discriminant Analysis and Logistic Regression under Multivariate Normality," Statistics in Biomedical, Public Health and Environmental Sciences, Sen, P.K. ed., Amsterdam: Elsevier (1985).
6. H. W. Chen, "A Case Study of Bank's Credit-Granting Model to Evaluate Individual Small-Amount Loans," thesis of Dept. of Finance, NKFUST, Taiwan (2002).
7. J. A. Ohlson, "Financial Ratios and the Probabilistic Prediction of Bankruptcy," *Journal of Accounting Research*, pp. 109-131 (1980).
8. J. Begley, J. Ming, and S. Watts, "Bankruptcy Classification Errors in the 1980s: An Empirical Analysis of Altman's and Ohlson's Models," *Review of Accounting Studies*, Vol. 1, No. 4, pp.267-284 (1996).
9. M. J. A. Berry and G. Linoff, "Mastering data mining: the art and science of customer relationship management," New York Chichester: Wiley Computer Publishing (2000).
10. M. C. Lin, "The Research on Evaluating the Risk of -Taking X Bank as an Example," thesis of Dept. of Business Management, TTU, Taiwan (2004).
11. T. H. Liang, J. B. Lin, and J. J. Liao, "The Best Oversampling Proportion and Variables Selection Procedure in the Predictive Model of Probability of Default of Personal Consumer Loans," *Information sciences 2007, Proceedings of the 10th Joint Conference* , pp. 383-389 (2007).
12. T. H. Liang and J. B. Lin, "Application Of Combined Variables In The Predictive Model Of Probability Of Default Of Personal Consumer Loans," *Journal of Data Analysis*, Vol. 3, No. 3, pp.135-150 (2008).
13. X. J. Shi, "Optimal Sample Pairing and Critical Value of Logistic Default Risk Modeling: The China Case," *Application of Statistics and Managemen*t, Vol. 25, No. 6, pp. 675-682 (2006).
14. Zavgren, "Assessing the vulnerability to failure of American industrial firms: a logistic analysis," *Journal of Business Finance and Accounting*, Vol. 12, pp.19-45 (1985).